

RESEARCH ARTICLE

Using machine learning to predict fire-ignition occurrences from lightning forecasts

Ruth Coughlan¹  | Francesca Di Giuseppe¹  | Claudia Vitolo¹  |
Christopher Barnard¹ | Philippe Lopez¹ | Matthias Drusch²

¹European Centre for Medium-range
Weather Forecasts (ECMWF),
Reading, UK

²European Space Agency-European Space
Research and Technology Centre (ESA-
ESTEC), Noordwijk, the Netherlands

Correspondence

Francesca Di Giuseppe, European Centre
for Medium-range Weather Forecasts
(ECMWF), Reading, UK.
Email: francesca.digiuseppe@ecmwf.int

Abstract

Lightning-caused wildfires are a significant contributor to burned areas, with lightning ignitions remaining one of the most unpredictable aspects of the fire environment. There is a clear connection between fuel moisture and the probability of ignition; however, the mechanisms are poorly understood and predictive methods are underdeveloped. Establishing a lightning–ignition relationship would be useful in developing a model that would complement early warning systems designed for fire control and prevention. A machine learning (ML) approach was used to define a predictive model for wildfire ignition based on lightning forecasts and environmental conditions. Three different binary classifiers were adopted: a decision tree, an AdaBoost and a Random Forest, showing promising results, with both ensemble methods (Random Forest and AdaBoost) exhibiting an out-of-sample accuracy of 78%. Data provided by a Western Australia wildfire database allowed a comprehensive verification on over 145 lightning-ignited wildfires in regions of Australia during 2016. This highlighted that in a minimum of 71% of the cases the ML models correctly predicted the occurrence of an ignition when a fire was actually initiated. The super-learner developed is planned to be used in an operational context to the enhance information connected to fire management.

KEYWORDS

AdaBoost, classifier, decision tree, lightning, lightning ignition, machine learning, Random Forest, wildfire

1 | INTRODUCTION

Although human activity is widely recognized as the primary cause of wildfires, in most ecosystems (Amatulli *et al.*, 2007; Archibald *et al.*, 2009; Ganteaume *et al.*, 2013), lightning-caused ignitions are not uncommon and the majority of burned areas in remote regions of the Boreal and Arctic biomes have a lightning origin (Krider *et al.*, 1980; Granström, 1993). Lightning is a major cause of

wildland fires in Canada (Hanes *et al.*, 2019), and in Alberta, for example, it is responsible for igniting 45% of reported wildfires and 71% of area burned (Blouin *et al.*, 2016). There is a clear connection between fuel dryness, its availability and the probability of a lightning ignition (Flannigan and Haar, 1986), even if complex non-environmental variables can play a role (Anderson, 2002). Simplified conceptual models to associate a danger of ignition from lightning have been proposed to derive generic

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. Meteorological Applications published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

danger levels. These can be used as one of various components to assess fire danger in rating systems (e.g. the lightning activity level in the National Fire Danger Rating System) (Fuquay, 1979). Still, the mechanisms associated with ignitions are poorly understood and predictive methods to identify a cause–effect relationship have been limited to specific regions (Flannigan and Wotton, 1991) and applied as interpretative tools for the analysis of occurred events (Wierzbowski *et al.*, 2002; Wotton and Martell, 2005).

In recent years, efforts have been devoted to the forecast of lightning activity using both physical models of various complexity (Price and Rind, 1992; Grewe *et al.*, 2001; MacGorman *et al.*, 2001; Allen and Pickering, 2002; McCaul Jr *et al.*, 2009; Lopez, 2016; Finney *et al.*, 2018) and statistical methods/machine learning (ML) approaches (Bates *et al.*, 2018; Mostajabi *et al.*, 2019). To a certain degree, lightning activity can be predicted with a reasonable level of confidence because it is linked to particular weather conditions. A key ingredient is intense convection: the rising of plumes of moist air in response to instability in the atmosphere, usually associated with the development of cumulonimbus clouds. Such atmospheric conditions often involve collisions between hydro-meteors such as graupel, hail, ice particles and liquid water droplets. The collisions may cause electric charges in those particles to separate, which leads to the creation of oppositely charged layers inside the clouds. The resulting electric fields can lead to the powerful discharges observed as lightning (Takahashi, 1978). One of the most significant challenges in assessing lightning's influence on ignition comes from the capability actually to predict whether a thunderstorm is going to be wet or dry, that is, accompanied by precipitation. Weather conditions and remotely sensed soil moisture data, as provided by microwave sensors, are ideally placed to discriminate between these two situations. Other environmental variables could also be either confounding or triggering factors. Ultimately, as the physical mechanisms to predict lightning ignitions are still largely unknown, the present paper explores the potential of ML techniques to provide a data-driven solution that could be applied globally. The aim is to derive a product able to link lightning predictions with the observed occurrence of fires and to provide a quantitative measure to identify those lightning flashes that could be conducive of fires.

As the availability of fire databases, including ignition causes, is limited, the study will not try to predict lightning–ignition events, but rather will focus on a methodology to identify the environmental factors occurring at the time when a fire was observed following lightning flashes predicted within the previous 24 hr. The result is to identify a forecasting product that could add valuable information to existing raw lightning forecasts in aid of early warning systems designed for fire control and prevention.

2 | DATA

2.1 | Lightning prediction

Using a parameterization developed by Lopez (2016), the European Centre for Medium-range Weather Forecasts (ECMWF) has been providing a lightning density forecast to its users since 2018. This is defined as the average number of flashes in a given time interval. Experiments show that the forecasts for lightning can have useful skill, at least, up to day 3. The skill depends on the averaging windows, so that when the integration is performed over the next 24 hr and for large regions, a very good agreement with observations can be achieved (Lopez, 2016). Lightning flashes can be categorized into either cloud-to-ground (CG) or intra- and intercloud (IC) discharges. Lopez's parameterization, in its operational implementation, predicts the total lightning flash (IC + CG) density (F_T) in (flashes·km^{−2}·day^{−1}) as:

$$F_T = \alpha Q_r \sqrt{\text{CAPE}} (\min\{Z_{\text{base}}, 1.8\})^2 \quad (1)$$

where α is a tuneable co-efficient (currently set to 37.5 to match the climatological annual global mean flash rate from the Tropical Rainfall Measuring Mission (TRMM) satellite Lightning Imaging Sensor (LIS)/Optical Transient Detector (OTD) climatology, as described by Cecil *et al.*, 2014); Z_{base} is the convective cloud base height (km); Q_r is a proxy for the charging rate resulting from the collisions between graupel particles and other types of hydrometeors; and CAPE is the convective available potential energy (J·kg^{−1}). An estimation of the cloud-to-ground density (F_{CG}) is calculated from the total flash density by a simple conversion:

$$a_1 = 0.021 \times H_c^4 \quad (2)$$

$$a_2 = 0.648 \times H_c^3 \quad (3)$$

$$a_3 = 7.49 \times H_c^2 \quad (4)$$

$$a_4 = 36.54 H_c - 64.09 \quad (5)$$

$$F(H_c) = \frac{1}{(a_1 - a_2 + a_3 - a_4)} \quad (6)$$

$$F_{CG} = F_T \times F(H_c) \quad (7)$$

where H_c is the thickness of the convective cloud above the freezing level; and $F(H_c)$ is derived using the partition

function proposed by Price and Rind (1993) applied to where $5.5 < H_c < 14$ km.

A database for CG lightning flash density (averaged over a 6 hr time window) was then computed using the TCo399 configuration of the ECMWF's Integrated Forecasting System (IFS) (roughly corresponding to a grid size of 28 km) for a full year in 2016.

2.2 | Lightning-ignition database

While there are local databases of wildfire ignitions categorized to have been initiated by lightning, they do not provide the global coverage needed for the present study. A training data set was achieved by identifying wildfire ignition points (IPs) and combining these data with the ECMWF's CG lightning density product to produce a data set of fire ignition points that could be potentially attributed to lightning (IPLs) (Figure 1).

The technique used to construct the IPL data set relied on the use of fire radiative power (FRP) ($\text{W}\cdot\text{m}^{-2}$) observations from the operational Global Fire Assimilation System (GFAS), which has been available on a daily basis since 2003 (Kaiser *et al.*, 2012; Di Giuseppe *et al.*, 2016a, 2016b). The procedure uses a series of decision nodes to identify new ignitions. We first exclude points that are burning as the result of previous fires (1) or are spreading from neighbouring areas (2). The procedure then removes fires that can be attributed to agriculture practices either by being shorter than a day or by being of low intensity ($\text{FRP} < 0.5 \text{ W}\cdot\text{m}^{-2}$) or if placed

in an agricultural biome (3). The agricultural biomes are defined as the combination of agricultural with organic soils (AGOS) and agricultural (AG) in the GFAS database (Van der Werf *et al.*, 2017).

Any newly active fires with no fire spread and not being agriculture related are deemed IPs. Depending on the CG lightning density, points identified as IPs (six) are then split into IPLs and non-IPLs. As non-IPLs observations are several orders of magnitude larger than IPLs, only a randomly selected sample (of the same sample size of IPLs) is used to train the binary classifier described below in Section 3.

2.3 | Environmental predictors

A comprehensive set of environmental variables is candidates that are possible predictors of lightning ignition. The selection is based on capturing not only the start of a fire event but also its sustainability. The probability of fuel ignition by lightning has been shown to be relatively independent of fuel moisture (Latham and Williams, 2001), while its survival is highly dependent on it (Wotton and Martell, 2005), thus the reason why the concept of dry lightning is of importance. To capture these conditions, a crucial aspect is physical consistency among the environmental factors and lightning prediction. Also given the global nature of the present study, and the focus on developing a predictive model that could be used in real time, the majority of the meteorological fields are taken from the IFS operational analysis as either raw outputs or derived parameters (Table 1).

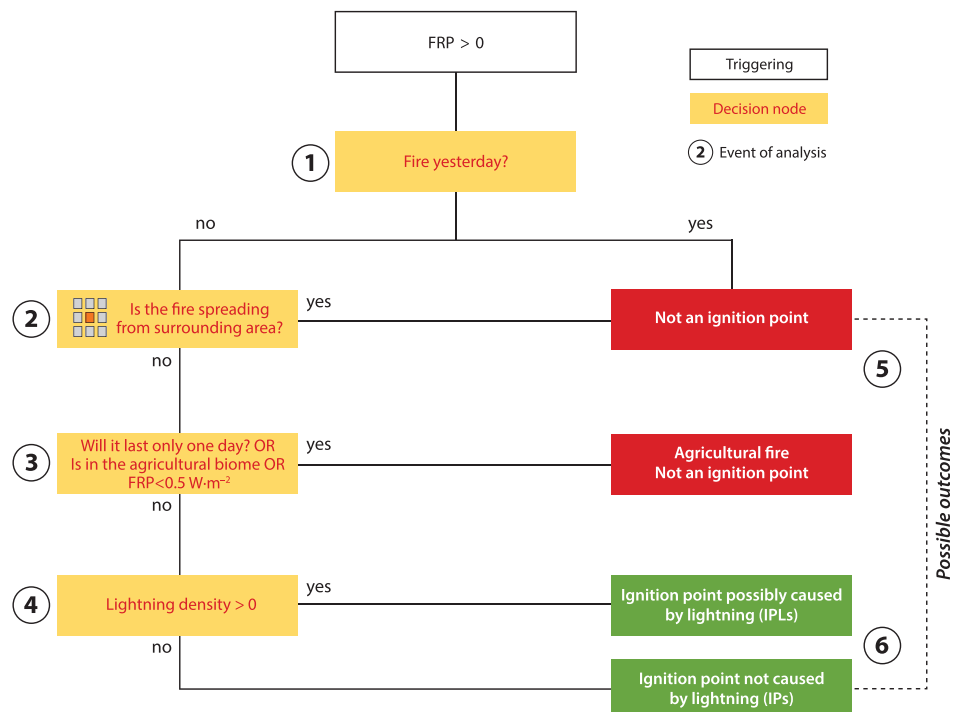


FIGURE 1 Flow chart representing the methodology employed in the construction of the data set that identifies fire ignition points possibly caused by lightning
Note: FRP, fire radiative power

One exception is provided by the soil moisture field. Even if great progress has been made in soil moisture initialization in weather prediction (Albergel *et al.*, 2012), this variable is still often used as a tuning parameter to adjust the prediction of screen-level diagnostics (Hess, 2001; Di Giuseppe *et al.*, 2011). Thus, despite recent efforts to include remote observations in soil moisture initialization in order to characterize better the water balance of the surface (De Rosnay *et al.*, 2013), the simplicity of the land surface schemes used in weather forecast models makes them inappropriate to represent fully hydrological processes across the globe (Zsoter *et al.*, 2019). For this reason remote observations of soil moisture are also tested as possible predictors. The soil moisture near real-time neural network product available from the Soil Moisture and Ocean Salinity (SMOS) satellite is available in near real time and used in the surface assimilation cycle of the ECMWF (Sabater *et al.*, 2017). Data are available in swath segments on the ISEA 4H9 grid, with grid data provided as discrete global grid (DGG) points, only available along the satellite track with a nominal resolution of 15 km. Soil moisture retrievals undergo a quality control and filtering identical to what was applied before ingestion in the ECMWF's IFS data assimilation. Data are discarded if any of these four conditions are met:

- DGG ID number < 0.
- Soil moisture < 0.
- Soil moisture uncertainty > 0.2 (m³·m⁻³).
- Radio frequency interference (RFI) probability > 20%.

Daily swath files are then regridded onto the 0.1 grid resolution using the inverse distance to a power

interpolation algorithm (GDAL/OGR contributors, 2020). Three days of daily swath files were missing between March 14 and 16, 2016; missing data were imputed using previous days observations. The availability time for these data varies because it depends on the availability of the satellite observations and the survival rate through the reprocessing and quality-assurance stages. However, usually data are available every 3–5 days.

As low fuel moisture, especially in the duff layer, is known to be an important agent for lightning ignition (Wotton and Martell, 2005), an additional parameter is considered which combines the drought code (DC) and the duff moisture code (DMC) of the Fire Weather Index model (Van Wagner, 1974, 1987). The DC averages moisture content of deep, compact organic layers of depth 10–20 cm with a long-term response (about 50 days) to weather variations. The DMC averages moisture content of loosely compacted organic layers of moderate depth 5–10 cm and is characterized by a medium-term response (about 10–12 days) to weather variations. A vegetation drought sensitivity index (VDI) was devised by Météo-France as a combination of DC and DMC and used in its operational practice (personal communication) to identify areas where fires can develop due to dry/hot and windy conditions (for the interpretation, see Table 2). The algorithm to calculate the VDI is publicly available (Vitolo *et al.*, 2017, 2018), and the ECMWF produces daily VDI layers as part of its experimental products.

Inevitably, the definition of the environmental parameters has a margin of discretion. Figure 2 highlights how lower level atmospheric parameters such as 2 m temperature, 2 m relative humidity (RH) and precipitation have a good discriminating power with the

ID	Parameter	Units	Variable tag
1	IFS volumetric soil water layer 1	(m ³ ·m ⁻³)	swvl1
2	IFS 2 m temperature	Kelvin	t2m
3	IFS 2 m relative humidity [▽]	%	RH
4	IFS soil type	Categorical (1–7)	slt
5	IFS low vegetation covfer	Fraction (0–1)	cvl
6	IFS high vegetation cover	Fraction (0–1)	cvh
7	IFS daily accumulated precipitation	mm·day ⁻¹	tp
8	IFS 5 days precipitation [▽]	mm·day ⁻¹	tp r5
9	IFS Avg 3 days soil moisture [▽]	(m ³ ·m ⁻³)	SM r3
10	IFS Avg 5 days soil moisture [▽]	(m ³ ·m ⁻³)	SM r5
11	IFS Avg 10 days soil moisture [▽]	(m ³ ·m ⁻³)	SM r10
12	IFS Avg 15 days soil moisture [▽]	(m ³ ·m ⁻³)	SM r15
13	IFS Avg 20 days soil moisture [▽]	(m ³ ·m ⁻³)	SM r20
14	SMOS NN soil moisture	(m ³ ·m ⁻³)	SM IDW
15	IFS vegetation drought index	Categorical (1–5)	VDI

TABLE 1 Parameters marked with an [▽] are calculated from the European Centre for Medium-range Weather Forecasts' (ECMWF) model Integrated Forecasting System (IFS) volumetric soil water layer 1 raw outputs. Numbers 4 and 15 are categorical variables, while numbers 5 and 6 are fractional masks

TABLE 2 Vegetation drought sensitivity index (VDI) numeric code, interpretation and corresponding drought code (DC) and duff moisture code (DMC) ranges

Numeric code	VDI interpretation	Conditions on DC and DMC
1	No fire vulnerability, corresponding to an important superficial humidification	$DMC \leq 20$ and $DC \leq 400$
2	Very limited drying. Small fires possible	$DMC \leq 20$ and $200 < DC \leq 700$ or $DMC \leq 50$ and $DC \leq 500$ or $DMC \leq 110$ and $DC \leq 350$ or $DMC \leq 170$ and $DC \leq 300$
3	Zone is considered vulnerable due to strong wind and low humidity	$DMC \leq 20$ and $DC > 700$ or $20 < DMC \leq 50$ and $500 < DC \leq 1,000$ or $50 < DMC \leq 70$ and $350 < DC \leq 750$ or $70 < DMC \leq 110$ and $350 < DC \leq 650$ or $110 < DMC \leq 170$ and $300 < DC \leq 500$ or $170 < DMC \leq 200$ and $DC \leq 450$ or $200 < DMC \leq 250$ and $DC \leq 300$
4	Important drying; the zone is considered vulnerable. Fires can occur in any conditions, except by high air moisture. Very severe fire weather danger by a moderate wind, even a low wind with a Foehn effect or very hot and very dry air	$20 < DMC \leq 50$ and $DC > 1,000$ or $50 < DMC \leq 70$ and $DC > 750$ or $70 < DMC \leq 110$ and $DC > 650$ or $110 < DMC \leq 170$ and $500 < DC \leq 1,000$ or $170 < DMC \leq 200$ and $450 < DC \leq 750$ or $200 < DMC \leq 250$ and $300 < DC \leq 650$ or $DMC > 250$ and $DC \leq 600$
5	Extreme drying; the zone is considered extremely vulnerable. Very big fire conditions are gathered. Permanent risk of very big fires on slope zones. Catastrophic fires are possible in any zones by moderate or strong wind	$110 < DMC \leq 170$ and $DC > 1,000$ or $170 < DMC \leq 200$ and $DC > 750$ or $200 < DMC \leq 250$ and $DC > 650$ or $DMC > 250$ and $DC > 600$

probability density function (PDF) for the IPLs shifted towards dryer and warmer conditions, as would be expected when fires ignite. Soil moisture also shows as a strong candidate as a predictor if using the IFS prediction or SMOS observations. The use of an averaging window does not seem to add additional information. Among the categorical parameters, the VDI suggests the non-IPL situation falling mostly into category 1 (no fire vulnerability), while the majority of the ignitions fall into categories 2–5, corresponding to higher levels of vulnerability.

2.4 | Validation data set

A large data set of lightning-ignition fires on lands throughout Western Australia managed by the Department of Biodiversity, Conservation and Attractions Western Australia was made available to the ECMWF through a scientific collaboration for the purposes of the study (for the geographical extension of the verification database, see Figure 3). The data collected use lookout towers and spotter aircraft with a manual cause-classification for fire events. The database reports several different attributes, including fire cause and fire discovery time. Even

though the detection network is used several times each day and the likelihood of a prolonged delay is small, there still could be a possible delay compared with the actual ignition time. In 2016, a total of 147 events attributed to lightning were observed.

3 | BINARY CLASSIFIERS FOR LIGHTNING-IGNITION PREDICTION

The study used supervised learning to find a mapping function between weather and environmental variables and a specific target. The target was a binary label identifying whether a fire event occurred and if it were caused by lightning. The ground truth used to train the model was, in turn, the result of the other model used to construct the data set of fire IPs possibly caused by lightning. We used models falling under the so-called binary classification algorithms allowing only two distinct class decisions (i.e. two possible outcomes): IPLs (corresponding to 1) or non-IPLs (corresponding to 0). To build a robust prediction system, three models of increasing complexity were tested: a single decision tree (DT); a Random Forest

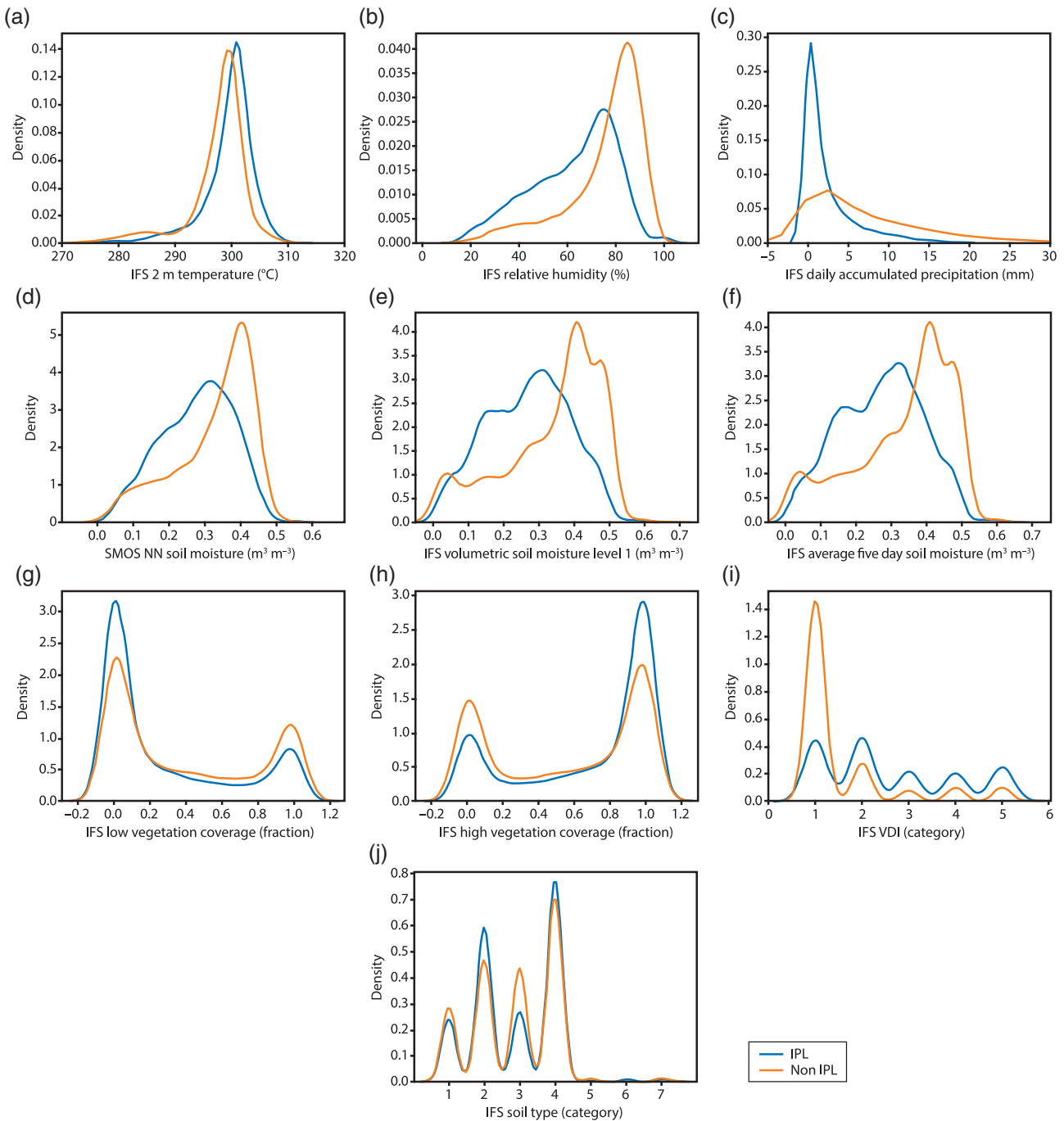


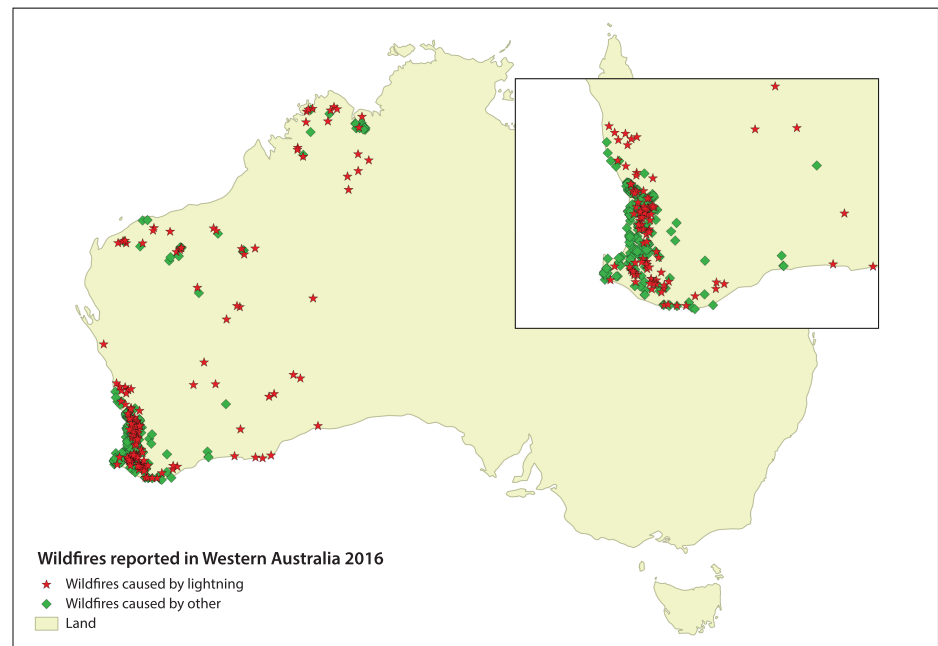
FIGURE 2 Comparison of probability distribution functions (PDFs) (as kernel density estimate) of fire ignition points possibly caused by lightning (IPL) and non-IPL observations in 2016 for the parameters: (a) Integrated Forecasting System (IFS) 2 m temperature; (b) IFS relative humidity (RH); (c) IFS daily accumulated precipitation; (d) Soil Moisture and Ocean Salinity (SMOS) neural network (NN) soil moisture; (e) IFS volumetric soil moisture level 1; (f) IFS average five day soil moisture; (g) IFS low vegetation coverage; (h) IFS high vegetation coverage; (i) IFS vegetation drought sensitivity index (VDI); and (j) IFS soil type

(RF) and an adaptive (Ada-) Boost (AD) applied to the DT. The reason for employing multiple models of increasing complexity is to explore the trade-off between the complexity of the ML approach and the accuracy of prediction.

Kotsiantis (2011) notes that DT techniques have been widely used to build classification models; such models

closely resemble human reasoning and are easy to understand. The goal of a DT is to create a set of rules (i.e. a tree) that minimize a loss function that measures the discrepancy between the predicted and the true values (Venkatasubramaniam *et al.*, 2017). The algorithm works by analysing a data set and constructing a set of rules,

FIGURE 3 Locations of ignition points for 2016 attributed to lightning and non-lightning causes. Database provided by the Department of Biodiversity, Conservation and Attractions Western Australia



which subsequently divides the data set into smaller and smaller subgroups until it can make a prediction that falls into one of the expected classes (Song and Lu, 2015). In general, decisions trees are very sensitive to the data on which they are trained. Thus, small changes to the training set can result in significantly different tree structures and, in turn, the predictions can be quite different. To overcome this limitation, an RF uses ensemble methods that combine several DT classifiers to produce better predictive performance than a single DT classifier. An added benefit is that several trees in an RF can be run in parallel, thus allowing a very large number of input variables to be handled and minimizing overfitting (Biau, 2012). Another strategy to improve a simple DT is provided by applying to it an adaptive boosting (AdaBoost). The AdaBoost algorithm trains sequentially a set of single-variable DTs (Ying *et al.*, 2012). It uses base learners of individual decision stumps (depth = 1), or small DTs, and trains them for several iterations, each trying to correct its predecessor. By doing so it “boosts” its predictive capability. This method, which can be applied to any classification algorithm, builds on top of a classifier rather than being a classifier by itself. Unlike other models, variable selection is pre-applied to the base learner rather than the ensemble learning algorithm.

The applied supervised learning uses data covering the whole of 2016, where all environmental parameters are available and CG lightning density is > 0 . This environmental conditions data set is then sieved into non-IPL class (labelled 0) and IPL class (labelled 1). Data are then divided into 85% training data and 15% testing data. The

same data sets are used with all the models in order to assess and compare performance accurately. There is a large disparity in sizes between the two classes with a significantly larger number of samples in the non-IPL class. This is known to lead to lower sensitivity when detecting the less populated class (Cheng *et al.*, 2015). In order to rebalance the training data set, fixed under-sampling (Lemaître *et al.*, 2017) was performed to generate a data set containing equally sized classes.

Models were built in two steps. First, a feature selection was performed combined with a 10-fold cross-validation (CV). An optimization procedure was then applied, referred to as hyperparameter tuning (or simply tuning), which allows the final predictive skill of the models to be improved. For details of these procedure and the parameter setting, see Appendix A1 in the supplemental data online.

4 | LIGHTNING MODELS

4.1 | Feature selection and model assessment

A very relevant aspect of the definition of the prediction model is the analysis of the selected predictors as a result of the applications of the recursive feature elimination (RFE) with CV. This step provides information on which physical processes have the largest importance when detecting lightning ignition-prone conditions (for the methodologies, see Appendix A1 in the supplemental data online).

TABLE 3 Summary table of the features used by each model, ordered by decreasing importance

Model	Features
Decision tree and AdaBoost	RH, t2m, tp r5, swvl1, SM r20, tp, cvh, VDI, SM r10, SM IDW
Random Forest	RH, t2m, tp r5, SM r20, swvl1, tp, cvh, SM r15, VDI, SM r3, cvl, SM r5, SM r10, slt, SM IDW

The best variable combination for the deterministic DT is achieved with 10 predictors, while all 15 predictors are deemed important by the RF (Table 3). The AdaBoost, a boosting method applied to the DT, relies on the variables selected by the base estimator. The prediction score for the DT tends to decrease with the inclusion of variables such as soil type, low-level vegetation cover, and three day, five day or 15 day average soil moistures. Looking at the PDF of soil type in Figure 2, for instance, it is evident that this variable does not discriminate well between IPLs and non-IPLs. Soil moisture, on the other hand, is shown to have a very good predictive skill. This is consistent across several averaging periods; therefore, the information becomes redundant when it is carried by several variables. For the ensemble RF, the use of all variables leads to the highest prediction score.

Each estimator is trained on the initial set of features and the importance of each feature is obtained through a “feature importance” attribute (Pedregosa *et al.*, 2011). This attribute provides a score that indicates how valuable each feature was during the construction of the model. The features listed in Table 3 are ordered based on the feature importance attribute; Figure 4 provides a quantitative comparison of the importance of each feature across the three models. For instance, for the AdaBoost model, the top three variables were volumetric soil water layer 1 (swvl1), 20 day average soil moisture (SM r20) and high-level vegetation cover (cvh). There is a high level of consistency between the DT and RF, with the two most important predictors being 2 m RH and 2 m temperature (t2m). The variables excluded from the DT are soil type, low-level vegetation cover and other soil moisture variables, which are generally ranked low in the RF.

Model accuracy can be improved by performing hyper-parameter optimization by looking at relevant elements of the model design (for details of the optimization performed, see Appendix A1 in the supplemental data online). A key aspect of optimization in the DT algorithms is pruning, which attempts to build a simpler

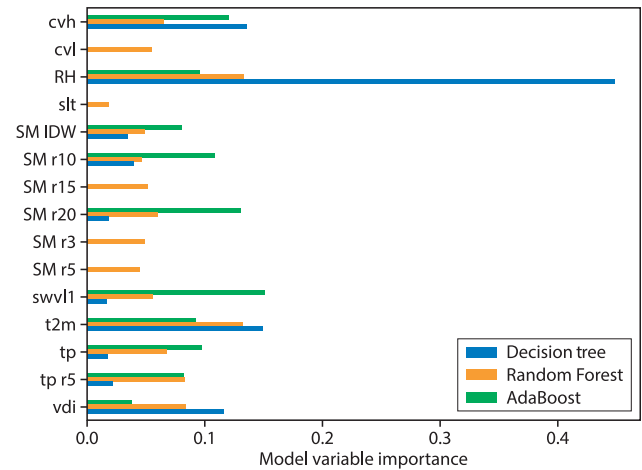


FIGURE 4 Model variable importance as a function of the three classifiers

model to help generalization and avoid overfitting to the training data. AdaBoost allows the further optimization of the DT identifying the optimum depth of the base trees. This is done empirically by assessing the performances of various depths (≤ 10) and comparing the CV mean test-fold score; the highest score is used to choose the optimum depth for the base estimator. The RF instead can be optimized by changing the number of the trees in the forest and the subsampling (bootstrapping) of the training data set for each tree in the forest. After tuning, predictive model performance can be assessed by using the independent test data set and the following metrics (Pedregosa *et al.*, 2011):

$$\begin{aligned}
 \text{Specificity} &= \frac{TN}{TN + FP} \\
 \text{Precision} &= \frac{TP}{TP + FP} \\
 \text{Sensitivity} &= \frac{TP}{TP + FN} \\
 \text{Balanced accuracy} &= \frac{1}{2} (\text{Sensitivity} + \text{Specificity}) \\
 F1 &= 2 \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}
 \end{aligned}$$

where the true positives (TP) are cases in which the classifier correctly predicted the IPLs; the true negatives (TN) are cases in which the classifier correctly predicted the non-IPLs; the false positives (FP) are cases in which the classifier predicted the IPLs, but the records were labelled as non-IPLs; and the false negatives (FN) are cases in which the classifier predicted the non-IPLs while they were labelled the IPLs.

The train accuracies, calculated using down-sampled/balanced data, were 76.00%, 78.11% and 78.13% for the

TABLE 4 Summary table comparing the results for the three classifiers, using the imbalanced test set

Model	True negatives (TN)	False positives (FP)	False negatives (FN)	True positives (TP)	Specificity	Precision	Sensitivity	Accuracy	F1 score
Decision tree	568,217	161,881	3,854	4,735	77.83	2.09	55.13	66.48	5.41
AdaBoost	564,419	165,679	2,698	5,891	77.31	2.72	68.59	72.95	6.54
Random Forest	576,275	153,823	3,024	5,565	78.93	2.67	64.79	71.86	6.63

DT, AB and RF, respectively. The test data set uses unbalanced data for which Table 4 reports the results from a randomly sampled month (i.e. April) and shows the four elements inside the contingency table (namely TP, TN, FP and FN), as well as: *specificity*, which is the TN rate; *precision*, which is the ability of the model not to label as positive a negative outcome; *sensitivity*, which is the ability of the model to detect all the positive outcomes and balanced accuracy score (a combination of sensitivity and specificity); and F1 score (a combination of sensitivity and precision). The use of balanced accuracy and F1 score avoids inflated performance estimates on imbalanced data sets (He and Ma, 2013).

As expected, the out-of-sample accuracy (calculated on the independent test data) is smaller than the in-sample accuracy (train data). The difference between in- and out-of-sample accuracies is relatively small, which suggests none of the models is overfitting. While the scores in Table 4 provide a useful indication of the expected accuracy of the model, the practical significance of the elements of the confusion matrix requires some additional comments. It must be stressed that the training data set collects all points where a *fire event* occurred, and the aim of the model is to attribute that ignition to a lightning strike, or not. This means that the model is trained to detect conditions that are prone to fire and *also* to attribute correctly the origin of the fire to lightning. Thus, the high number of FP (and consequently low F1 score) should be read having this perspective in mind. The model can predict correctly that a fire occurred but incorrectly tag it as generated by a lightning event. This might be due to meteorological conditions that do not allow one to discern between lightning and other causes of fires, as well as to a possibly incorrect lightning forecast. Thus, while from a ML perspective an FP is an erroneous result, in the present application this is expected due to the uncertainties associated with how the IPL data set was built, that is, by associating a cause–effect relationship with lightning and fires when occurring at the same time.

In summary, while we report all the scores for completeness, we stress that the FP result has not the same relevance as a FN when a lightning ignition occurred but

was not picked up by the model. The FNs could lead to serious consequences in the use of this model in an operational context. A more detailed analysis of the implications these verification scores have in a real context is performed in the next session where observed fires are analysed.

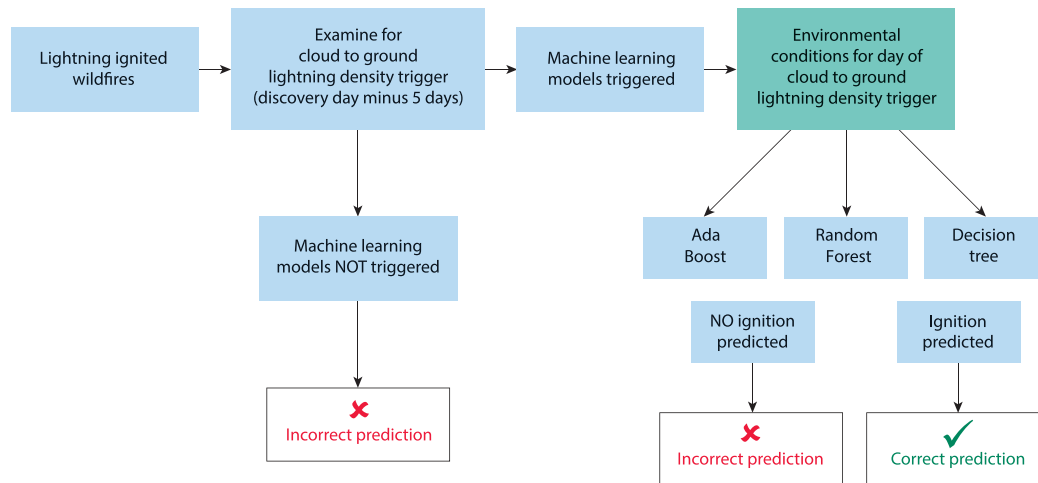
4.2 | Lightning-ignition prediction accuracy in observed fire events

Several factors are important to consider in the verification of the ML algorithms in order to forecast a lightning-ignition event. Possibly the most important is if the forecast of the CG lightning is accurate enough actually to trigger the predictive model in the first instance. The second aspect is the instant at which to calculate predictions from the models, because there can be a delay between the time of a lightning ignition and the date of discovery (Dowdy and Mills, 2012). To address these two aspects, we adopted strategies to understand the relative impacts of both conditions. Following the schematic depicted in Figure 5, the verification addresses three perspectives:

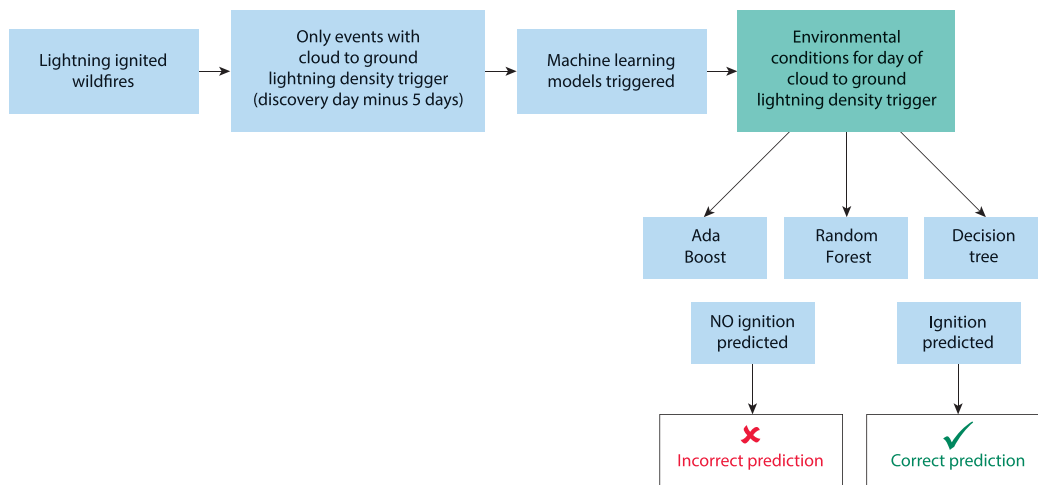
- If the models are correctly triggered by the CG lightning density forecast, looking at the discovery day and a time window of five days before this date, what is the success of both the CG lightning density product combined with each individual model?
- Only correctly considering the predicted events by the CG lightning forecast, what is the individual success of the predictions of each model?
- If the CG lightning forecast had been perfect and triggered all the wildfire events on the discovery date (DD), what would the predictive success be for each individual model?

In order to assess the success of predictions, we used the sensitivity score, as defined in Section 4.1. If we consider verification perspective 1, that is, the overall capability of the system to predict a lightning ignition, we find that 104 of the 147 lightning events coincided with

Verification perspective 1



Verification perspective 2



Verification perspective 3

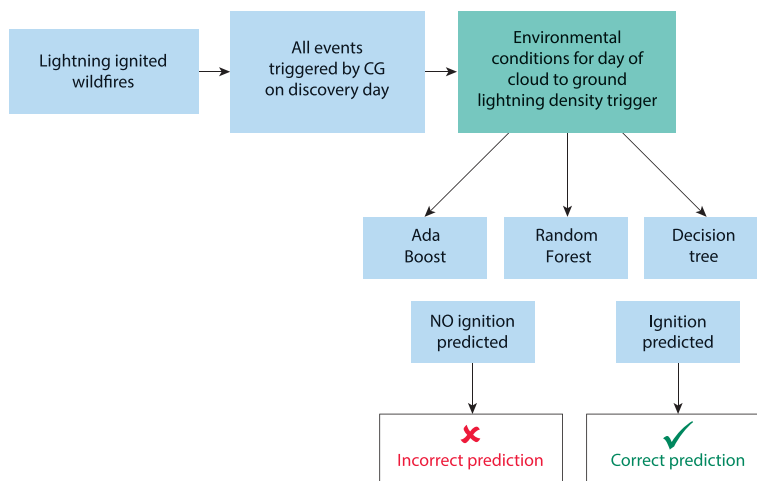


FIGURE 5 Methodology for the verification of lightning-caused wildfire events highlighting the three perspectives of verification implemented. Perspective 1 examines the overall capability of the system to identify lightning ignition as the combined success of the lightning forecast and the machine learning (ML) models. Perspective 2 assesses the capability of every ML model to predict lightning ignition when the lightning forecast is correct. Perspective 3 overrides the lightning forecast and assesses the capability of the three ML models to predict lightning ignition in all cases

TABLE 5 Number of lightning trigger misses (false negatives, FN) considering up to five days wide temporal window before the discovery date (DD) and sensitivity for the decision tree (DT) model, Random Forest (RF) model and AdaBoost (AD) model for verification perspective 1 (Ver 1) and verification perspective 2 (Ver 2)

Day of discovery	FN	Sensitivity (Ver 1)			Sensitivity (Ver 2)		
		DT	RF	AD	DT	RF	AD
DD	43	0.55	0.55	0.52	0.78	0.78	0.73
DD-1	25	0.65	0.63	0.60	0.78	0.76	0.72
DD-2	19	0.68	0.64	0.61	0.78	0.73	0.70
DD-3	15	0.68	0.64	0.62	0.76	0.71	0.69
DD-4	13	0.68	0.64	0.63	0.76	0.71	0.69
DD-5	11	0.70	0.64	0.63	0.76	0.71	0.69

TABLE 6 Sensitivity for the three models assuming a perfect lightning forecast on the discovery day

Model	True positives (TP)	False negatives (FN)	Sensitivity
Decision tree	110	37	0.75
Random Forest	111	36	0.76
AdaBoost	104	43	0.71

lightning forecast on the same day when the fire was discovered. This increases to 136 if we also consider lightning forecasts occurring in any of the previous five days from the DD, with the largest contribution (18 events) given by the day before the discovery date (DD - 1). For the combined success of the CG lightning density product with the three models' predictions, see Table 5. The results shows that, by only considering the DD, 43 events are not correctly predicted by the lightning forecast. However, by extending the search to the days before the DD, we reduce this to only 11 undetected events. Looking at the sensitivity of the models in Table 5, there is also an increase in the score as the number of detected events increases. While the increased number of lightning "triggerings" is an obvious consequence widening the temporal search for a lightning event, the sensitivity increase also arises from the capability of the models to use environmental conditions prone to fire. If we only consider events that were correctly predicted by lightning and the model performance of the classifiers (Table 5), the sensitivity increases initially, with a decrease in the score as the number of detected events increases. This decrease is due to ambiguity between the day of discovery and the "trigger" day environmental conditions.

Lastly, assuming a perfect lightning forecast able always to predict a strike when it happened, what would be the predictive capability of the models? In this third perspective the sensitivities of all three models are between 0.71 and 0.76 when using the discovery day as the ignition day. Degrading performance is recorded if a temporal delay is used (data not shown). Overall, the three models have an accuracy in the range of 71–76%. This is a good performance rate due to its position

between a random unbiased binary classifier and the perfect binary classifier, which have by definition accuracies of 50% and 100%, respectively. Interestingly, the best predictions in this case are provided by the RF model (Table 6). Performance differences between models are limited, with variances occurring due to the contrasting approach between the various models, the different variables used and their importance, and how each model appoints a prediction result.

5 | CONCLUSIONS

The aim of the study was to develop a lightning-ignition model. The analysis performed established a definable lightning-ignition relationship by only looking at three of the more popular machine learning (ML) methods. We acknowledge that further improvements could be made by using other ensemble methods, for example, a gradient boosting classifier, which is similar to AdaBoost, but identifies the shortcomings of the prediction not by weights but by a loss function. Although an empirical approach was adopted to identify predictors that were assumed to be most relevant in a lightning-ignition relationship, further variables (such as wind speed and solar radiation) could be added to explore whether supplementary variables add extra benefit. Further, soil moisture variables used for the study were based on the top layer (0–7 cm). Laboratory studies have shown that some soils report burn depths much greater than this range (Benscoter *et al.*, 2011), which could suggest that the inclusion of an extra feature such as the depth profile of the soil moisture might be worthwhile. Lastly, the

inclusion of a predictor related to fuel availability beyond the climatological vegetation cover adopted here could be beneficial. Still, the results are encouraging as for most of the real cases all models can correctly point out those environmental conditions favourable to an ignition. This shows how empirical models can have great potential to overcome some of the limitations of process-based models. It also suggests that the future will highly benefit from the adoption of a pragmatic approach that combines traditional model formulation with artificial intelligence methods for model parameter estimation and process optimization (Reichstein *et al.*, 2019).

For the purpose of developing an operational product from the analysis, the most interesting information comes from the analysis of the missed events because it has been shown that the skill in lightning prediction plays a significant role. If lightning is not correctly predicted, the ML models will fall into the non-IPL class (where IPL is fire ignition points that could be potentially attributed to lightning) regardless of whether or not the environmental conditions were prone to a lightning ignition. Moreover, the definition of the data set of the IPLs used to train the binary classifiers assumes that the cloud-to-ground (CG) lightning density product is accurate in space and time, along with the removal of ignition points (IPs) of agricultural nature and short duration, which leaves the potential to miss smaller lightning-ignited fires that are suppressed by rain or other activities. Thus, the quality of the lightning forecast affects not only the triggering of the model but also the definition of the classification. However, there are clear advantages in using a forecast model: (1) the forecast has global coverage while observations are spatially sparse; (2) there is a dynamic consistency and spatial coherence with the other environmental variables; and (3) there is an opportunity to extend the prediction into the future. However, it is clear from the analysis performed that a system that accounts for the inherent stochasticity of the prediction has to be put in place to deliver a product that could be useful in complementing early warning systems designed for fire control and prevention. Weather forecast errors are characterized by nonlinear growth, meaning that a small inaccuracy can lead to very different physical states (Lorenz, 2000). Accounting for these errors is traditionally done through the use of ensemble prediction systems where several simulations are performed starting from slightly different initial conditions and model configurations (Molteni *et al.*, 1996; Buizza *et al.*, 1999). An ensemble system means that the forecast can be interpreted as probabilistic rather than deterministic, and uncertainties can then be assigned to the final prediction, and this can boost confidence in the decision process during emergency situations as a cost-loss analysis can also be associated with the different scenarios (Clove *et al.*, 2017).

The use of an ensemble system that is very relevant for this application also implies a substantial increase in the computational time and could jeopardize the feasibility of delivery products in real time. From a preliminary test performed at the European Centre for Medium-range Weather Forecasts (ECMWF) supercomputer facility, we estimated that with the model run time to < 10 min for an ensemble member and the possibility of running parallel tasks, even a probabilistic product could be delivered with a delay of around 1 hr from when the lightning forecast becomes available.

Although the model was designed to recognize the IPs due to lightning from other causes of ignitions, further work will test the ability of the model also to classify non IPs as if they were non-IPLs. This goes in the direction of making the model work in an operational setting and benefit decision-making for fire management and prevention. In this context, land use/cover and the temporal evolution of fire radiative power at a location and its neighbouring area would become active features of the model. Also, a probabilistic outcome, in terms of the probability of ignition due to lightning, would suit tools for decision-making better than the current deterministic outcome.

5.1 | Steps to build a binary classifier model

The construction of the binary classifiers in the present study has four steps. The first consists of splitting the initial data set into training (85%) and testing (15%) sets. In this stage, fixed under-sampling (Lemaître *et al.*, 2017) was performed on the training portion to generate a data set containing equally sized classes. The training set was used to train the model through cross-validation (CV), while the testing set was used to assess the out-of-sample accuracy.

The second step is feature (variable) selection, that is, which combinations of the 15 environmental variables are more able to predict the conditions present in the training data set, which are identified using a combination of recursive feature elimination (RFE) and CV. The RFE is built by recursively considering smaller and smaller sets of variables. First, the estimator is trained on the initial set of variables, and the importance of each variable is obtained through a “feature importance” attribute (Pedregosa *et al.*, 2011) that attributes importance related to how a variable was used during the construction of the model. Essentially, the more a variable is used to make key decisions, the higher its relative importance. The least important variables are removed from the current set of variables and the procedure is recursively

repeated on the reduced set of features until the desired number of features to select is reached. The CV is combined with the RFE to identify the best set of variables. On each recursion, the CV is performed and the recursion with the highest CV score is deemed the optimum set of variables for a model.

The third step is the definition of a CV methodology. Normally when developing an ML model, data are split into training and test sets, with the training set used to learn/train the model and the test set to assess the model. The test set is used to appraise the model's performance by using evaluation metrics, such as accuracy. This method, however, can be unreliable as the accuracy obtained for one test set can differ to the accuracy obtained for another. The CV provides a solution to this: it evaluates and compares a model by splitting data into segments—one is used to train a model and the other is used to validate (the interactions of the various partitions replace the need for a separate “validation” set). A K -fold CV is the basic form of the CV where train data are divided into k equally (or nearly equally) sized segments or folds. Subsequently, k iterations of training and validation are performed, and within each iteration a different fold is held out for validation while the remaining $k - 1$ -fold are used for learning (Refaeilzadeh *et al.*, 2009). For the present study, the number of folds chosen was 10, with the overall mean CV test fold score taken as the assessment metric for models.

The final step is hyper-parameter tuning. A hyper-parameter is a parameter whose value is used to control the learning process (Feurer and Hutter, 2019). Thus, hyper-parameter tuning is the problem of choosing a set of optimal hyper-parameters for a learning algorithm (Ghawi and Pfeffer, 2019). To obtain good accuracy in a model, the easiest way is to test and compare different parameter combinations (DeCastro-García *et al.*, 2019), and this is how hyper-parameter tuning is approached. Hyper-

parameter tuning for the study was undertaken by creating different unique models with different sets of hyper-parameters using tuning tools from the Python scikit-learn package. Within this package, we used the GridSearchCV function for both decision tree (DT) and AdaBoost models and the RandomizedSearchCV and GridsearchCV functions for Random Forest (RF). Model assessment was undertaken by using the highest mean CV test fold score, with the hyper-parameters of this model considered the optimum for the learning process. (Pedregosa *et al.*, 2011).

TABLE 8 Parameters used to create the parameter grid for the Random Forest RandomizedSearchCV, with optimal parameters for a learning algorithm highlighted in bold

Grey parameter	Parameter values for grid
N estimators	100, 200, 300, 400, 500, 600, 700, 800 , 900, 1,000
Bootstrap	True , false
Warm start	True , false
Criterion	Gini , entropy
Maximum depth	3, 5, 8, 10, 15, 20 , 25, 30, none
Minimum samples split	5 , 10, 15, 20, 30, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1
Minimum samples leaf	3, 4 , 6, 8, 10, 15, 20, 25, 30, 0.01, 0.02, 0.03, 0.04, 0.05
Maximum features	Auto , Sqrt, Log2, none
Maximum leaf nodes	5, 10, 15, 20, 25, 30, 40, none

TABLE 7 Parameters used to create the parameter grid for the GridSearchCV for the decision tree with optimal parameters for a learning algorithm highlighted in bold

Grey parameter	Parameter values for grid
Criterion	Gini , entropy
Splitter	Best , random
Maximum depth	3, 5, 8 , 10, 15, 20, none
Minimum samples split	5, 10, 15, 20, 30, 0.01 , 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1
Minimum samples leaf	3, 4, 6, 8, 10, 15, 20, 25, 30 , 0.01, 0.02, 0.03, 0.04, 0.05
Maximum features	Auto, Sqrt, Log2, none
Maximum leaf nodes	5, 10, 15, 20, 25, 30, 40, none

TABLE 9 Optimum depth results for the base estimator for AdaBoost, with the best depth highlighted in bold

Grey depth of base estimator	Score
1	0.7564
2	0.7512
3	0.7391
4	0.7257
5	0.7161
6	0.7213
7	0.74278
8	0.7477
9	0.7624
10	0.7684

TABLE 10 Parameters used to create the parameter grid for the AdaBoost GridSearchCV, with optimal parameters for a learning algorithm highlighted in bold

Grey parameter	Parameter values for grid
N estimators	100, 200, 300, 400, 500, 600, 700 , 800, 900, 1,000
Learning rate	0.001, 0.01, 0.1

In the spirit of reproducibility, Tables 9–12 provide the main parameter settings for all three classifiers used in hyper-parameter optimization. In particular, Tables 7 and 10 define the parameter grid for the GridSearchCV of the DT and AdaBoost, respectively. Similarly, Table 8 defines the parameter grid for the RF RandomizedSearchCV. Finally, Table 9 shows the results of the optimum depth assessments for the base estimator for AdaBoost.

ACKNOWLEDGEMENTS

The authors thank the Department of Biodiversity, Conservation and Attractions Western Australia for providing the verification data, with special thanks to Lachie McCaw (Department of Biodiversity, Conservation and Attractions Western Australia) and Adam Leavesley (Australian Capital Territory Parks and Conservation Service) for facilitating the data acquisition. The work was conducted through a collaboration between the European Centre for Medium-range Weather Forecasts (ECMWF) and European Space Agency (ESA) (contract number 4000125399/18/I-BG). The authors declare no competing interests. Ruth Coughlan designed the experiments, performed the analysis, generated the graphical outputs and wrote part of the material. Francesca Di Giuseppe co-designed the experiments and wrote the paper. Claudia Vitolo and Christopher Barnard contributed to the design of the experiments. Philippe Lopez contributed with the lightning forecast database. All authors revised the paper.

ORCID

Ruth Coughlan  <https://orcid.org/0000-0001-5293-0048>

Francesca Di Giuseppe  <https://orcid.org/0000-0001-9829-0429>

Claudia Vitolo  <https://orcid.org/0000-0002-4252-1176>

REFERENCES

- Albergel, C., De Rosnay, P., Balsamo, G., Isaksen, I. and Muñoz-Sabater, J. (2012) Soil moisture analyses at ECMWF: evaluation using global ground-based in situ observations. *Journal of Hydrometeorology*, 13(5), 1442–1460.
- Allen, D.J. and Pickering, K.E. (2002) Evaluation of lightning flash rate parameterizations for use in a global chemical transport model. *Journal of Geophysical Research: Atmospheres*, 107(D23), ACH 15–1–ACH 15–21. <https://doi.org/10.1029/2002JD002066>.
- Amatulli, G., Pérez-Cabello, F. and de la Riva, J. (2007) Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. *Ecological Modelling*, 200(3–4), 321–333.
- Anderson, K. (2002) A model to predict lightning-caused fire occurrences. *International Journal of Wildland Fire*, 11(4), 163–172.
- Archibald, S., Roy, D.P., van Wilgen, B.W. and Scholes, R.J. (2009) What limits fire? An examination of drivers of burnt area in Southern Africa. *Global Change Biology*, 15(3), 613–630.
- Bates, B.C., Dowdy, A.J. and Chandler, R.E. (2018) Lightning prediction for Australia using multivariate analyses of large-scale atmospheric variables. *Journal of Applied Meteorology and Climatology*, 57(3), 525–534.
- Benscoter, B.W., Thompson, D.K., Waddington, J.M., Flannigan, M. D., Wotton, B.M., de Groot, W.J. and Turetsky, M.R. (2011) Interactive effects of vegetation, soil moisture and bulk density on depth of burning of thick organic soils. *International Journal of Wildland Fire*, 20(3), 418–429.
- Biau, G. (2012) Analysis of Random Forest model. *Journal of Machine Learning Research*, 13(38), 1063–1095.
- Blouin, K.D., Flannigan, M.D., Wang, X. and Kochtubajda, B. (2016) Ensemble lightning prediction models for the province of Alberta, Canada. *International Journal of Wildland Fire*, 25 (4), 421–432.
- Buizza, R., Milleer, M. and Palmer, T.N. (1999) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560), 2887–2908.
- Cecil, D.J., Buechler, D.E. and Blakeslee, R.J. (2014) Gridded lightning climatology from TRMM-LIS and OTD: dataset description. *Atmospheric Research*, 135–136, 404–414. <https://doi.org/10.1016/j.atmosres.2012.06.028>.
- Cheng, F., Yang, K. and Zhang, L. (2015) A structural SVM based approach for binary classification under class imbalance. *Mathematical Problems in Engineering*, 2015, 269856. <https://doi.org/10.1155/2015/269856>.
- Cloke, H.L., Pappenberger, F., Smith, P.J. and Wetterhall, F. (2017) How do I know if I've improved my continental scale flood early warning system? *Environmental Research Letters*, 12(4), 044006.
- De Rosnay, P., Drusch, M., Vasiljevic, D., Balsamo, G., Albergel, C. and Isaksen, L. (2013) A simplified extended Kalman filter for the global operational soil moisture analysis at ECMWF. *Quarterly Journal of the Royal Meteorological Society*, 139(674), 1199–1213.
- DeCastro-García, N., Muñoz Castañeda, Á.L., Escudero García, D. and Carriegos, M.V. (2019) Effect of the sampling of a dataset in the hyperparameter optimization phase over the efficiency of a machine learning algorithm. *Complexity*, 2019, 6278908. <https://doi.org/10.1155/2019/6278908>.
- Di Giuseppe, F., Cesari, D. and Bonafé, G. (2011) Soil initialization strategy for use in limited-area weather prediction systems. *Monthly Weather Review*, 139(6), 1844–1860.
- Di Giuseppe, F., Pappenberger, F., Wetterhall, F., Krzeminski, B., Camia, A., Libertá, G. and Miguel, J.S. (2016a) The potential predictability of fire danger provided by numerical weather prediction. *Journal of Applied Meteorology and Climatology*, 55(11), 2469–2491.

- Di Giuseppe, F., Rémy, S., Pappenberger, F. and Wetterhall, F. (2017) Improving Forecasts of Biomass Burning Emissions with the Fire Weather Index. *Journal of Applied Meteorology and Climatology*, 56(10), 2789–2799. Available at: <https://journals.ametsoc.org/view/journals/apme/56/10/jamc-d-16-0405.1.xml>. Accessed Jan 4, 2021.
- Dowdy, A. and Mills, G. (2012) Characteristics of lightning-attributed fires in south-east Australia. *International Journal of Wildland Fire*, 21, 521–524. <https://doi.org/10.1071/WF10145>.
- Feurer, M. and Hutter, F. (2019) Hyperparameter optimization. In: Hutter, F., Kotthoff, L. and Vanschoren, J. (Eds.) *Automated Machine Learning. The Springer Series on Challenges in Machine Learning*. Cham: Springer.
- Finney, D.L., Doherty, R.M., Wild, O., Stevenson, D.S., MacKenzie, I. A. and Blyth, A.M. (2018) A projected decrease in lightning under climate change. *Nature Climate Change*, 8(3), 210–213.
- Flannigan, M.D. and Wotton, B.M. (1991) Lightning-ignited forest fires in northwestern Ontario. *Canadian Journal of Forest Research*, 21(3), 277–287.
- Flannigan, M.D. and Haar, T.H.V. (1986) Forest fire monitoring using NOAA satellite AVHRR. *Canadian Journal of Forest Research*, 16(5), 975–982.
- Fuquay, D.M. (1979) *A Model for Predicting Lightning Fire Ignition in Wildland Fuels*, Vol. 217 Washington, DC: Inter-mountain Forest and Range Experiment Station, Forest Service, US Department of Agriculture.
- Ganteaume, A., Camia, A., Jappiot, M., San-Miguel-Ayán, J., Long-Fournel, M. and Lampin, C. (2013) A review of the main driving factors of forest fire ignition over Europe. *Environmental Management*, 51(3), 651–662.
- GDAL/OGR contributors (2020). *GDAL/OGR Geospatial Data Abstraction software Library—Inverse Distance to a Power*. Open Source Geospatial Foundation. Available at: https://gdal.org/tutorials/gdal_grid_tut.html#inverse-distance-to-a-power. Accessed Jul 12, 2019.
- Ghawi, R. and Pfeffer, J. (2019) Efficient hyperparameter tuning with grid search for text categorization using KNN approach with BM25 similarity. *Open Computer Science*, 9, 160–180. <https://doi.org/10.1515/comp-2019-0011>.
- Granström, A. (1993) Spatial and temporal variation in lightning ignitions in Sweden. *Journal of Vegetation Science*, 40(6), 737–744.
- Grewe, V., Brunner, D., Dameris, M., Grenfell, J.L., Hein, R., Shindell, D. and Staehelin, J. (2001) Origin and variability of upper tropospheric nitrogen oxides and ozone at northern mid-latitudes. *Atmospheric Environment*, 35(20), 3421–3433. [https://doi.org/10.1016/S1352-2310\(01\)00134-0](https://doi.org/10.1016/S1352-2310(01)00134-0).
- Hanes, C.C., Wang, X., Jain, P., Parisien, M.-A., Little, J.M. and Flannigan, M.D. (2019) Fire-regime changes in Canada over the last half century. *Canadian Journal of Forest Research*, 49(3), 256–269.
- He, H. and Ma, Y. (2013) *Imbalanced Learning: Foundations, Algorithms, and Applications*. New York, NY: John Wiley & Sons, Ltd.
- Hess, R. (2001) Assimilation of screen-level observations by variational soil moisture analysis. *Meteorology and Atmospheric Physics*, 77(1–4), 145–154.
- Kaiser, J.W., Heil, A., Andreae, M.O., Benedetti, A., Chubarova, N., Jones, L., Morcrette, J.J., Razinger, M., Schultz, M.G., Suttie, M. and Van der Werf, G.R. (2012) Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power, 9, 1, 527–554.
- Kotsiantis, S.B. (2011) Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261–283. <https://doi.org/10.1007/s10462-011-9272-4>.
- Krider, E.P., Noggle, R.C., Pifer, A.E. and Vance, D.L. (1980) Lightning direction-finding systems for forest fire detection. *Bulletin of the American Meteorological Society*, 61(9), 980–986.
- Latham, D. and Williams, E. (2001) *Lightning and Forest Fires*. Amsterdam, Netherlands: Elsevier.
- Lemaître, G., Nogueira, F. and Aridas, C.K. (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(1), 559–563.
- Lopez, P. (2016) A lightning parameterization for the ECMWF integrated forecasting system. *Monthly Weather Review*, 144 (9), 3057–3075. <https://doi.org/10.1175/MWR-D-16-0026.1>.
- Lorenz, E. (2000) The butterfly effect. *World Scientific Series on Nonlinear Science Series A*, 39, 91–94.
- MacGorman, D.R., Straka, J.M. and Ziegler, C.L. (2001) A lightning parameterization for numerical cloud models. *Journal of Applied Meteorology*, 40(3), 459–478.
- McCaul, E.W., Jr., Goodman, S.J., LaCasse, K.M. and Cecil, D.J. (2009) Forecasting lightning threat using cloud-resolving model simulations. *Weather and Forecasting*, 24(3), 709–729.
- Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122 (529), 73–119.
- Mostajabi, A., Finney, D.L., Rubinstein, M. and Rachidi, F. (2019) Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *npj Climate and Atmospheric Science*, 2(1), 1–15.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Price, C. and Rind, D. (1993) What determines the cloud-to-ground lightning fraction in thunderstorms. *Geophysical Research Letters*, 20, 463–466. <https://doi.org/10.1029/93GL00226>.
- Price, C. and Rind, D. (1992) A simple lightning parameterization for calculating global lightning distributions. *Journal of Geophysical Research: Atmospheres*, 97(D9), 9919–9933.
- Refaeilzadeh, P., Tang, L. and Liu, H. (2009) Cross-validation. In: Liu, L. and Özsu, M.T. (Eds.) *Encyclopedia of Database Systems*. Boston, MA: Springer.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. and Carvalhais, N. (2019) Deep learning and process understanding for data-driven earth system science. *Nature*, 566 (7743), 195–204.
- Sabater, J.M., Rodriguez-Fernandez, N., Sabater, J., Richaume, P., de Rosnay, P., Kerr, Y., Albergel, C., Drusch, M. and Mecklenburg, S. (2017) SMOS near-real-time soil moisture product: processor overview and first validation results. *Hydrology and Earth System Sciences*, 21(10), 5201–5216.

- Song, Y.Y. and Lu, Y. (2015) Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130–135. <https://doi.org/10.11919/j.issn.1002-0829.215044>.
- Takahashi, T. (1978) Riming electrification as a charge generation mechanism in thunderstorms. *Journal of the Atmospheric Sciences*, 35(8), 1536–1548.
- Van der Werf, G.R., Randerson, J.T., Giglio, L., Van Leeuwen, T.T., Chen, Y., Rogers, B.M., Mu, M., Van Marle, M.J.E., Morton, D.C., Collatz, G.J., et al. (2017) Global fire emissions estimates during 1997–2016. *Earth System Science Data*, 9(2), 697–720.
- Van Wagner, C.E. (1974) *Structure of the Canadian Forest Fire Weather Index*. Alberta, Canada: Environment Canada, Forestry Service.
- Van Wagner, C.E. (1987) *Development and Structure of the Canadian Forest Fire Weather Index System*, Vol. 35. Ottawa: Canadian Forestry Service, Headquarters.
- Venkatasubramaniam, A., Wolfson, J. and Mitchell, N. (2017) Decision trees in epidemiological research. *Emerging Themes in Epidemiology*, 14, 11. <https://doi.org/10.1186/s12982-017-0064-4>.
- Claudia Vitolo, Francesca Di Giuseppe, and Mirko D'Andrea (2017). Caliver r package version 1.0. *Software release*. doi: <https://doi.org/10.5281/zenodo.376613>. Available at: <https://github.com/ecmwf/caliver>.
- Vitolo, C., Di Giuseppe, F. and D'Andrea, M. (2018) Caliver: an r package for CALibration and VERification of forest fire gridded model outputs. *PLoS One*, 13(1), e0189419. <https://doi.org/10.1371/journal.pone.0189419>.
- Wierzchowski, J., Heathcott, M. and Flannigan, M.D. (2002) Lightning and lightning fire, central cordillera, Canada. *International Journal of Wildland Fire*, 11(1), 41–51.
- Wotton, B.M. and Martell, D.L. (2005) A lightning fire occurrence model for Ontario. *Canadian Journal of Forest Research*, 35(6), 1389–1401.
- Ying, C., Qi-Guang, M., Jia-Chen, L. and Lin, J. (2012) Advance and prospects of Adaboost algorithm. *Acta Automatica Sinica*, 39(6), 745–758.
- Zsoter, E., Cloke, H., Stephens, E., de Rosnay, P., Muñoz-Sabater, J., Prudhomme, C. and Pappenberger, F. (2019) How well do operational numerical weather prediction configurations represent hydrology? *Journal of Hydrometeorology*, 20(8), 1533–1552.

How to cite this article: Coughlan R, Di Giuseppe F, Vitolo C, Barnard C, Lopez P, Drusch M. Using machine learning to predict fire-ignition occurrences from lightning forecasts. *Meteorol Appl.* 2021;28:e1973. <https://doi.org/10.1002/met.1973>